¿En que sentido el aprendizaje supervisado es posible?

Curso Aprendizaje Automático Aplicado

Julio Waissman

Recapitulando

Decimos que $f \approx h^*$ ssi

$$E_i(h^*) \approx 0$$

y

$$E_o(h^*) \approx E_i(h^*)$$

$E_i(h^*) \approx 0$

- Problema de optimización
- Encontrar h^* equivale a encontrar el vector de parámetros θ^* tal que

$$\theta^* = \arg\min_{\theta \in \Theta} \frac{1}{M} \sum_{i=1}^{M} loss(y^{(i)}, h_\theta(x^{(i)}))$$

$$E_o(h^*) \approx E_i(h^*)$$

- Generalización
- Diferencia entre aprendizaje y optimización
- Vamos a usar una noción que parece una broma:

Aprendizaje Probablemente Aproximadamente Correcto (PAC Learning)

$$E_o(h^*) \approx E_i(h^*)$$

Hoy vamos a dedicarnos a ver en que sentido es posible que un modelo ajustado por aprendizaje supervisado *generalice*:

Desigualdad de Hoeffding

$$\Pr[|E_o(h^*) - E_i(h^*)| \geq \epsilon] \leq 2\exp(-2\epsilon^2 M)$$

donde M es el número de datos y ϵ la diferencia entre el error en muestra y el error fuera de muestra impuesto.

Entonces, el planteamiento $E_o(h^*) \approx E_i(h^*)$ es PAC

¿Algún problema con la desigualdad de Hoeffding?

- Si lanzo una moneda 10 veces, ¿Cual es la probabilidad de obtener águila las 10 veces?
- Si 1000 personas lanzan una moneda 10 veces, ¿Cual es la probabilidad que *alguna* de las personas obtengan águila las 10 veces?

¿Esto que significa?

- Supongamos un problema de clasificación binaria con 10 instancias en el conjunto de entrenamiento.
- *Algoritmo de aprendizaje*: clasificar en forma aleatoria.
- ¿Cual es la probabilidad de clasificar bien las 10 instancias?
- ¿Cual es la probabilidad de clasificar bien las 10 instancias en *alguna* iteración, si el algoritmo se entrena con un máximo de 1000 *epoch*?

Traduciendo

$$\Pr[|E_o(h^*) - E_i(h^*)| \geq \epsilon] \leq \Pr[\bigcup_{h \in \mathcal{H}} |E_o(h) - E_i(h)| \geq \epsilon]$$

$$\Pr[|E_o(h^*) - E_i(h^*)| \geq \epsilon] \leq \sum_{h \in \mathcal{H}} \Pr[|E_o(h) - E_i(h)| \geq \epsilon]$$

Y el problema de aprendizaje queda como...

$$\Pr[|E_o(h^*) - E_i(h^*)| \geq \epsilon] \leq 2N \exp(-2\epsilon^2 M)$$

donde N es el número de hipótesis posibles en el conjunto $\mathcal{H}$.

¿Entonces no es posible el aprendizaje?

Tranquilos, esta es una *cota superior* muy superior, vamos a tratar de hacerla más chiquita.

- Vamos a bosquejar el problema de generalización sólo para la clasificación binaria
- El procedimiento se puede generalizar a regresión pero ya se usa otra caja de herramientas en matemáticas que se sale de los alcances de este curso.

Clasificación binaria

- $h_\theta : \mathcal{X} \to \{-1, 1\}, \quad h_\theta \in \mathcal{H}$
- Una gran cantidad de translapes entre diferentes hipótesis
- Respecto al conjunto de aprendizaje, muchas hipótesis son iguales

Dicotomías

- Hipótesis $h : \mathcal{X} \to \{-1, 1\}, \quad h_\theta \in \mathcal{H}$
- Dicotomía
  $h : \{x^{(1)}, \ldots, x^{(M)}\} \to \{-1, 1\}, \quad h_\theta \in \mathcal{H}(x^{(1)}, \ldots, x^{(M)})$
- $|\mathcal{H}| = N$, muy seguramente infinito
- $|\mathcal{H}(x^{(1)}, \ldots, x^{(M)})| \leq 2^M$

La función de crecimiento

$$m_{\mathcal{H}}(M) = \max_{x^{(1)},\ldots,x^{(M)} \in \mathcal{X}} |\mathcal{H}(x^{(1)}, \ldots, x^{(\mathcal{M})})|$$

- Acotado a $m_{\mathcal{H}}(M) \leq 2^M$

Un poco mejor

$$\Pr[|E_o(h^*) - E_i(h^*)| \geq \epsilon] \leq 2m_{\mathcal{H}}(M)\exp(-2\epsilon^2 M)$$

$$2m_{\mathcal{H}}(M)\exp(-2\epsilon^2 M) \leq 2 \cdot 2^M \exp(-2\epsilon^2 M)$$

pero todavía no lo suficiente, necesitamos más ya que $m_{\mathcal{H}}(M)$ puede ser exponencial en M.

Vamos a acotar $m_{\mathcal{H}}(M)$

- ¿Que significa $m_{\mathcal{H}}(M)$?
- Ejemplos:
  - Rayos positivos
  - Intervalos positivos
  - Conjuntos convexos
  - Clasificación lineal

¿Y cuando es posible acotar $m_{\mathcal{H}}(M)$?

- Si existe un valor d tal que $m_{\mathcal{H}}(d) < 2^d$
- Entonces, para todo $M > d, m_{\mathcal{H}}(M) < 2^M$
- Si d no es infinito, entonces puede haber esperanza de aprender con $\mathcal{H}$

Pero no está todavía probado que el aprendizaje sea posible.

La dimensión VC $d_{VC}(\mathcal{H})$

- Ese valor d se llama **dimensión VC** de $\mathcal{H}$, $d_{VC}(\mathcal{H}) = d$
- $d_{VC}(\mathcal{H})$ es el mayor número de puntos que pueden ser clasificados *sin error* por hipótesis en $\mathcal{H}$
- Si no existe tal valor, $d_{VC}(\mathcal{H}) = \infty$ y no hay esperanza de aprender con $\mathcal{H}$

Vamos a llamar $k = d_{VC}(\mathcal{H}) + 1$ el *valor de ruptura* de $\mathcal{H}$

La función $B(M, k)$

$B(M, k)$ es el *máximo número de dicotomías* que pueden generar M puntos con *cualquier modelo* $\mathcal{H}$ con valor de ruptura k

- Si $M < k$, $B(M, k) = 2^M$
- Si $k = 1$, $B(M, k) = 1$
- Si $M = 1$, $B(M, k) = 2$, para $k \geq 2$

¿Y esto con que fin?

Pues lo que vamos a tratar de probar es lo siguiente:

$$\Pr[|E_o(h^*) - E_i(h^*)| \geq \epsilon] \leq 2m_{\mathcal{H}}(M)\exp(-2\epsilon^2 M)$$
$$2m_{\mathcal{H}}(M)\exp(-2\epsilon^2 M) \leq 2 \cdot B(M,k)\exp(-2\epsilon^2 M)$$
$$2B(M,k)\exp(-2\epsilon^2 M) \leq 2 \cdot f(M,k)\exp(-2\epsilon^2 M)$$
$$2f(M,k)\exp(-2\epsilon^2 M) \leq 2 \cdot 2^M \exp(-2\epsilon^2 M)$$

Y si logramos demostrar que $f(M,k)$ sea polinomial en M, entonces **el aprendizaje es posible**.

Acotando $B(M, k)$

Sea $B(M, k) < 2^M$ con el conjunto $\mathcal{D} = \{x^{(1)}, \ldots, x^{(M)}\} \subset \mathcal{X}$

- Sean los conjuntos S', S^+ y S^-, una partición de las asignaciones de $\mathcal{D}$ que pueden ser bien clasificados.
- S^+ y S^- tienen asignaciones con la misma clasificación para los primeros $M - 1$, pero $x^{(M)}$ es clasificado como $+1$ y -1 respectivamente.
- S' tiene el resto de las asignaciones.

Acotando $B(M, k)$

- Sean $\alpha = |S'|$ y $\beta = |S^+| = |S^-|$
- Entonces,

$$B(M, k) = |S'| + |S^+| + |S^-| = \alpha + 2\beta$$

Acotando $B(M, k)$

- Si quito $x^{(M)}$ de $\mathcal{D}$, *al menos* las asignaciones en S^+ y S' siguen siendo clasificables por $\mathcal{H}$ en los primeros $M-1$ puntos.

$$\alpha + \beta \leq B(M-1, k)$$

- Si, además, considero ahora una $\mathcal{H}$ con valor de ruptura $k-1$, *al menos* las asignaciones en S^+, al ser iguales que en S^- salvo $x^{(M)}$ siguen siendo clasificables.

$$\beta \leq B(M-1, k-1)$$

Encontrando una función f(M, k)

$$B(M,k) \leq B(M-1,k) + B(M-1,k-1)$$

Si $f(M,k) = f(M-1,k) + f(M-1,k-1)$ entonces

$$B(M,k) \leq f(M,k)$$

Vamos a proponer:

$$f(M,k) = \sum_{i=0}^{k-1} \binom{M}{i}$$

Encontrando una función f(M, k)

$$f(M,k) = \sum_{i=0}^{k-1} \binom{M}{i}$$

- Si $k = 1$, $f(M,k) = 1$
- Si $M = 1$, $f(M,k) = 2$ para $k \geq 2$
- Satisface la relación de recurrencia
  $f(M,k) = f(M-1,k) + f(M-1,k-1)$

$f(M,k)$ es una función acotada por M^{k-1} ($\mathcal{O}(M^{d_{VC}(\mathcal{H})})$)

El aprendizaje supervisado si es posible

Si $d_{VC}(\mathcal{H})$ finito, entonces $m_{\mathcal{H}}(M)$ es $\mathcal{O}(M^{d_{VC}})$

$$\Pr[|E_o(h^*) - E_i(h^*)| \geq \epsilon] \leq 4 m_{\mathcal{H}}(2M) \exp(-\frac{1}{8}\epsilon^2 M)$$

La desigualdad de Vapnik--Chervonenkis

¿Como calcular la d_{VC}

- Una posibilidad es con los *grados de libertad*
- No es un calculo correcto, pero es una aproximación que suele ser adecuada
- Cuidado con el conjunto $\mathcal{H}$

¿Y cuantos datos se necesitan para que el aprendizaje exista?

Simplificando la desigualdad VC por la funcion de crecimiento

$$\Pr[|E_o(h^*) - E_i(h^*)| \geq \epsilon] \leq \delta$$

$$\delta \text{ es } O\left(M^{d_{VC}} e^{-M}\right)$$

- Vamos a graficar esa simplificación

La regla de oro para la generalización

$$M \geq 10 d_{VC}(\mathcal{H})$$

- ¿Siempre aplica?